Unsupervised Machine Translation

Список литературы


Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato.
Unsupervised machine translation using monolingual corpora only.
arXiv preprint arXiv:1711.00043, 2017.


Alex Graves and Jurgen Schmidhuber.
Framewise phoneme classification with bidirectional lstm and other neural network architectures.
Neural Networks, 18(5-6):602–610, 2005.


Yunsu Kim and Jiahui Geng Hermann Ney.
Improving unsupervised word-by-word translation with language model and denoising autoencoder.
2018.


Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

Постановка задачи

Рассматривается задача построения модели перевода текста без использования параллельных текстов, т.е. пар одинаковых предложений на разных языках.
Пример.

- Что это?
- Вы говорите на украинском языке?

- Що це?
- Ви говорите українською мовою?

Данная задача возникает при построении моделей перевода для низкоресурсных языков (т.е. языков, для которых данных в открытом доступе немного).

Описание эксперимента

Для украинско-русской пары:

- Непараллельная выборка: мультиязычные статьи из Wikipedia.
- Параллельный корпус: OpenSubtitles-2018.

Для французско-английской пары:

- Параллельный корпус: multi30k.
- Метрика измерения качества: BLEU.

Обучение проводилось на обеих выборках, валидация на параллельном корпусе.

Результаты эксперименотов

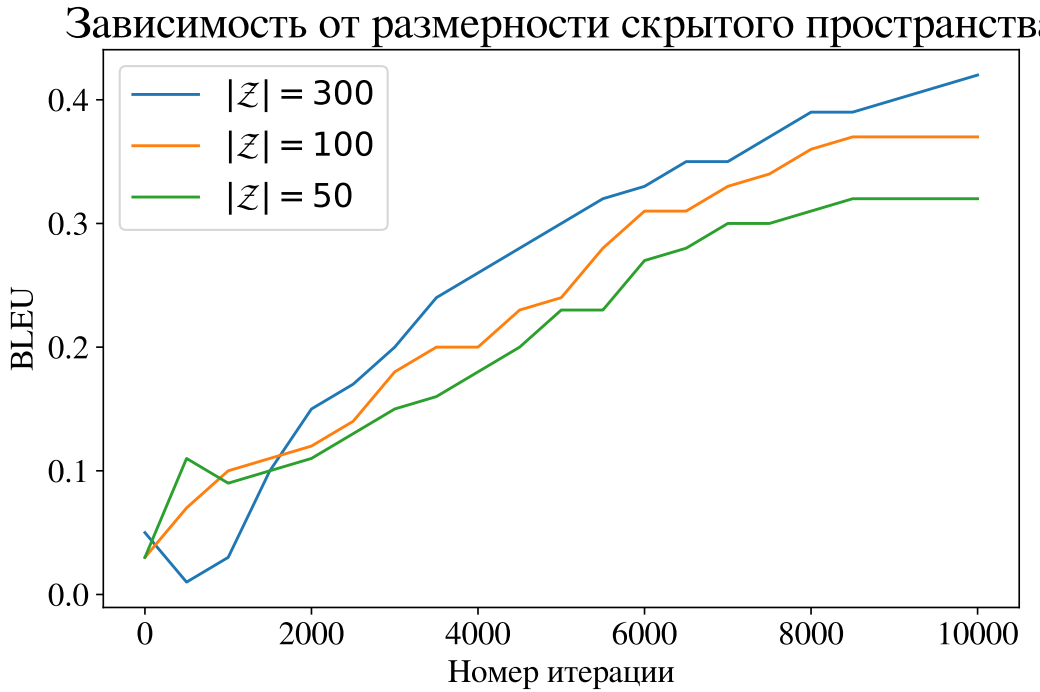

Figure: Зависимость BLEU от размерности скрытого пространства.

Результаты эксперименотов

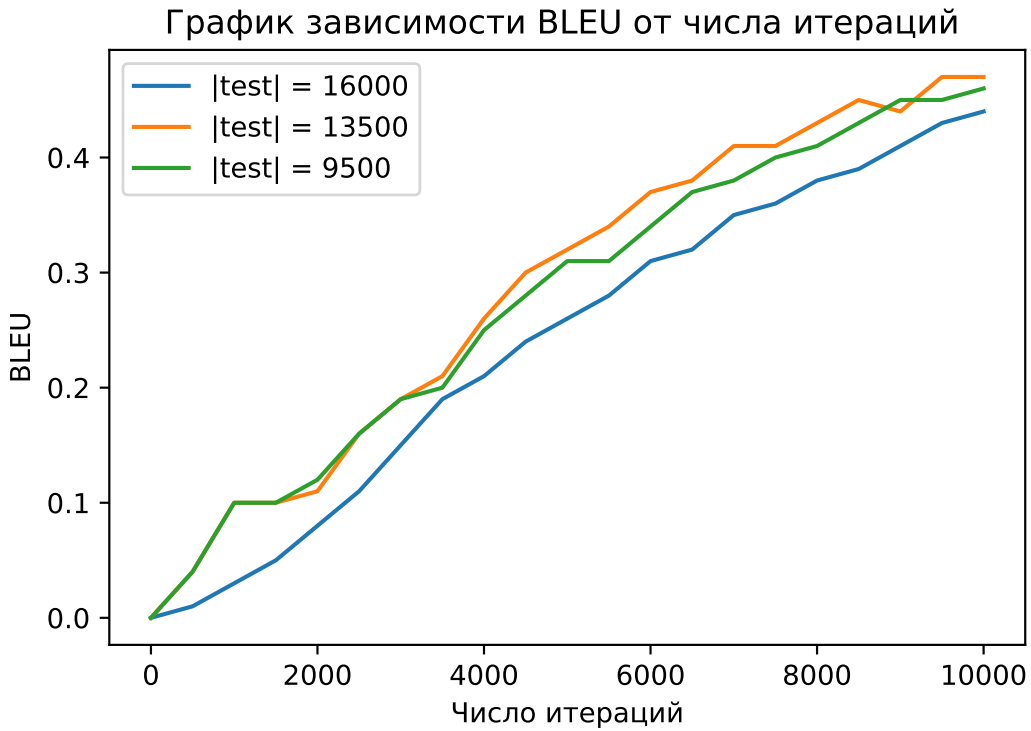

Figure: Зависимость BLEU от размера обучающей выборки.

Перевод с русского на украинский язык и обратно

Проблема: в языках слова могут иметь очень много форм. Из-за этого процесс обучения крайне долгий, вид кривых обучения не меняется кардинально при изменении модели.

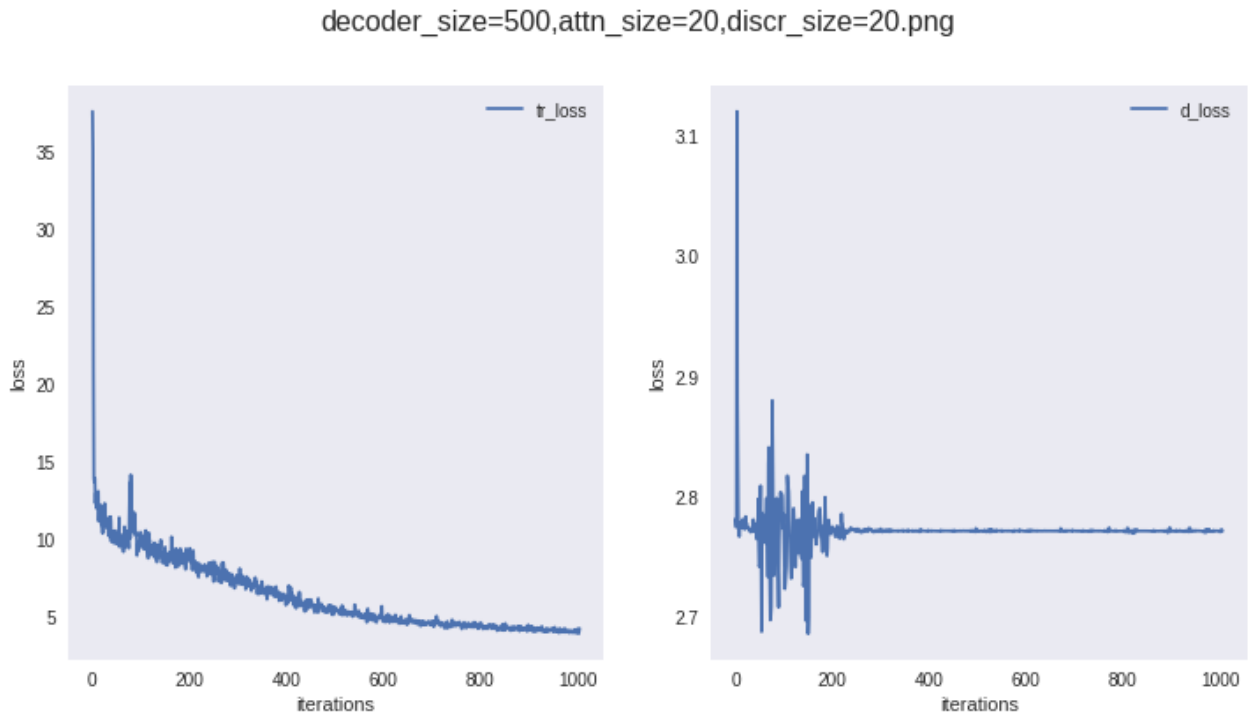

Figure: Типичный вид кривых обучения на выборке русский-украинский

Некоторое решение - Стемминг!

Стемминг — процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова.

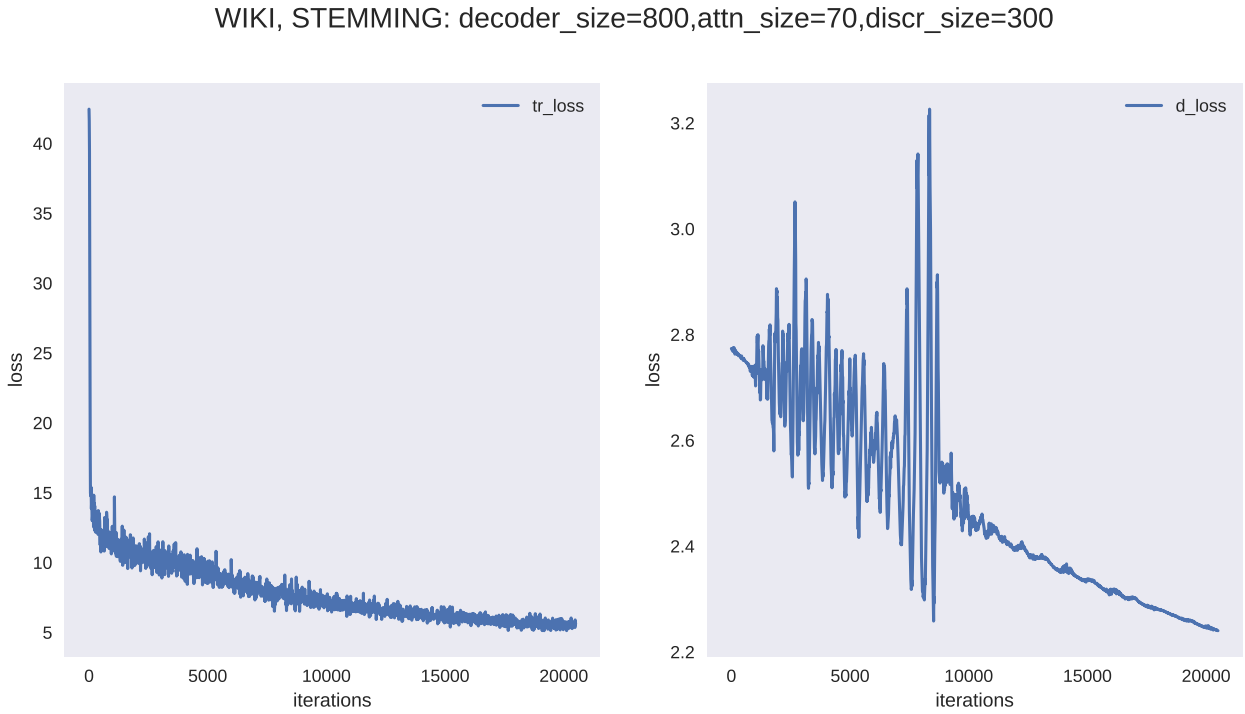

Figure: Кривые обучения в случае, когда всем словам были удалены последние две буквы

Некоторое решение - Стемминг!

Table: Результаты, достигнутые для русско-украинского перевода.

Модель(decoder,discriminator,attention)	BLEU после 10^5 итераций
Простая: $100, 100, 50$	0.113
Промежуточная: $200, 300, 50$	0.219
Промежуточная: $800, 600, 150$	0.222
Сложная: $2000, 2000, 150$	0.293
Промежуточная: $800, 300, 70$ + STM	0.413
Сложная: $2000, 2000, 150$ + STM	0.427

Выводы

- Качество найденных русско-украинских выборок не является достаточным как для построения хорошей модели, так и для валидации
- Была найдена одна из возможных проблем, связанная со структурой языков

Дальнейшие планы:

-
-